Originally published in: Calzolari, Nicoletta/Kan, Min-Yen/Hoste, Veronique/Lenci, Alessandro/Sakti, Sakriani/Xue, Nianwen (Eds.): The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Paris: European Language Resources Association (ELRA), 2024. Pp. 12553-12563.

Out of the mouths of MPs: Speaker Attribution in Parliamentary Debates

Ines Rehbein♠, Josef Ruppenhofer♡, Annelen Brunner♣, Simone Ponzetto♠
Universität Mannheim♠, FernUniversität Hagen♡, Leibniz-Institut der Deutschen Sprache Mannheim♣

ines.rehbein@uni-mannheim.de, josef.ruppenhofer@fernuni-hagen.de,
brunner@ids-mannheim.de, simone.ponzetto@uni-mannheim.de

Abstract

This paper presents GePaDe$_{SpkAtt}$, a new corpus for speaker attribution in German parliamentary debates, with more than 7,700 manually annotated events of speech, thought and writing. Our role inventory includes the sources, addressees, messages and topics of the speech event and also two additional roles, medium and evidence. We report baseline results for the automatic prediction of speech events and their roles, with high scores for both, event triggers and roles. Then we apply our model to predict speech events in 20 years of parliamentary debates and investigate the use of factives in the rhetoric of MPs.

Keywords: speaker attribution, factives, political text analysis

1. Introduction

Identifying who says what to whom is an essential prerequisite for analysing human communication. The complexity of the task, however, is often underestimated by assuming that the words produced by the speaker only reflect their own point of view. This, however, is far from true, as illustrated in Figure 1. The figure shows an excerpt from a parliamentary debate of the German Bundestag where the speaker switches back and forth between presenting their own view and reporting and citing other speakers. Thus, when analysing text, it is crucial to identify the correct source for each speech event. Furthermore, studying how speakers construct their own arguments relative to the views of other speakers, either to back up their own claim or to attack the others' perspective, is an intruiging research topic in itself.

In order to investigate these questions, we need annotated resources that allow us to train models that learn to identify speech events in unstructured text, together with their respective speakers, messages and addressees. This paper presents GePaDe$_{SpkAtt}$, a novel resource for speaker attribution in parliamentary debates from the German Bundestag.[1] The new corpus contains 267 speeches given by 195 different speakers, with a total of more than 200,000 tokens. The data includes around 7,700 annotated speech events with more than 11,000 annotated roles.

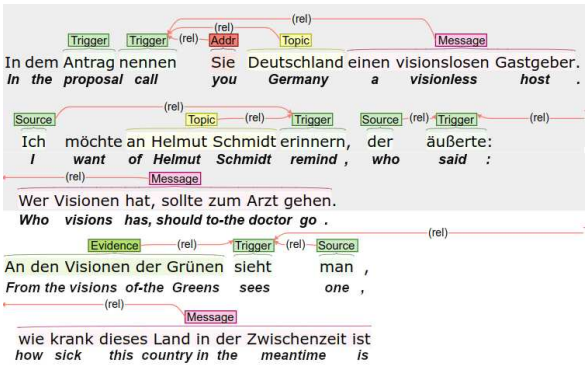

Figure 1: Example for speaker attribution in parliamentary debates.

Next, we adapt a transformer-based Semantic Role Labelling system to predict speaker attribution in parliamentary debates, and present baseline results on our new data set. The system obtains an accuracy of 95% for identifying speech events in the development and test set, for role prediction, we achieve an F1 score of around 80%. We then use our speaker attribution system to identify *speech* events (e.g., say, whisper, discussion), *thought* (e.g., think, assume, idea) and *writing* (e.g., write, scribble, application) in a large corpus of German parliamentary debates, spanning over a range of 20 years (2002–2023), and investigate the use of factive predicates in the rhetoric of MPs. Our results show that politicians make significantly more use of factives when they are in government than when their party is in opposition.

The paper is structured as follows. We review related work on speaker attribution in §2. The creation of our new resource is described in §3.

[1] We follow previous work and use the term *speaker attribution* to refer to the task of identifying speech events in text and attributing them to their respective sources. Please note that the source of the speech event is often different from the person who gave the speech, as illustrated in Example 1.

LREC-COLING 2024, pages 12553–12563
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0
Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: https://nbn-resolving.org/urn:nbn:de:bsz:mh39-127145

Section 4 explains our experimental setup and validation of our speaker attribution system. In §5, we apply our system to a large corpus of parliamentary debates and explore the use of factives in the German Bundestag. Finally, we conclude and outline avenues for future research in §6.

2. Related Work

2.1. Speaker attribution

Much recent work has been devoted to quote detection, mostly with the goal of extracting information from newswire text (Pouliquen et al., 2007; Krestel et al., 2008; Pareti et al., 2013; Pareti, 2015; Scheible et al., 2016). Other related work comes from the field of opinion mining and has targeted the identification of opinion holders (speakers) and the targets of the opinions (Choi et al., 2005; Wiegand and Klakow, 2012; Johansson and Moschitti, 2013).

Many studies have addressed speaker attribution in novels and other literary works, in the context of computational literary studies. Elson and McKeown (2010) were among the first to propose a supervised machine learning model for quote attribution in literary text. He et al. (2013) extend their supervised approach by including contextual knowledge from unsupervised actor-topic models. Almeida et al. (2014) and Fertmann (2016) combine the task of speaker identification with coreference resolution. Grishina and Stede (2017) test the projection of coreference annotations, a task related to speaker attribution, using multiple source languages. Muzny et al. (2017) improve on previous work on quote and speaker attribution by providing a cleaned-up dataset, the QuoteLi3 corpus, which includes more annotations than the previous datasets. They also present a two-step deterministic sieve model for speaker attribution on the entity level and report a high precision for their approach.[2] Papay and Padó (2020) annotate direct and indirect quotations in 19th century English literature while Kim and Klinger (2018) extend the speaker attribution task to capture emotion trigger phrases and the experiencers, targets and causes of the emotion.

While many studies have addressed the task of quote detection or speaker attribution in English text from the literary domain or in news articles, less work has been done for other languages and genres. Krug et al. (2018) focus on German literary text and release the he DROC corpus which includes around 2,000 manually annotated quotes and annotations for speakers and their mentions in 90 fragments from German literary prose. The Corpus REDEWIEDERGABE of Brunner et al. (2020) is substantially larger and presents a German-language historical corpus of literary texts and non-fiction (historical newspapers and magazines), with detailed annotations for speech, thought and writing. Dönicke et al. (2022) address a task related to speaker attribution, i.e., identifying whether a certain text passage is written from the perspective of the narrator of the novel or from the author's point of view, or whether it reflects the view of a character in the novel. Interestingly, they show that including annotator bias in the model can improve results.

Less work has been done for other domains. A noteworthy exception is Ruppenhofer et al. (2010) who present preliminary work on speaker attribution in text from the political domain, using German cabinet protocols. Also related is the work of Ruppenhofer et al. (2016) on the extraction of subjective expressions and their sources and targets in political speeches from the Swiss parliament. As our focus is on analysing the language of political debates, we extend the work of Ruppenhofer et al. (2010) and create a new, manually annotated resource for speaker attribution with around 13,000 clauses and more than 200,000 tokens.

Brunner et al. (2020) was an important basis for our annotation in that we take into account not only speech events, but also thought and writing. The annotation scheme shares several ideas with Brunner's work but has a somewhat different label inventory inspired by work in Automatic Semantic Role Labeling in the FrameNet mode (Baker et al., 1998). We describe the creation of these resources in the next section.

3. Data and Annotation

We present a new dataset for speaker attribution in data from the political domain, specifically, parliamentary debates from the German Bundestag. GePaDe$_{SpkAtt}$ includes manually annotated cues that trigger events of speech, writing and thought.[3] In addition, we annotate the arguments of the trigger, including the SOURCE, ADDRESSEE, MESSAGE, MEDIUM, TOPIC and EVIDENCE for the speech event. Table 1 shows examples for the different categories in our schema. We now describe our data, annotation setup and annotation procedure.

Data The data includes debates from the German Bundestag, retrieved from Deutscher Bun-

[2] When optimised for precision, the system obtains a score >95% on the development set from *Pride and Prejudice*.

[3] In the reminder of the paper, we use the term "speech event" to refer not just to speech events but also to events of thought and writing.

Cue/Role name	Description	Example
CUE	the cue that evokes the STW event	Merkel spoke$_{Cue}$ to the people.
SOURCE	Source of the STW event	Merkel$_{Source}$ spoke to the people.
MEDIUM	Medium of the STW event	The constitution$_{Medium}$ states ...
MESSAGE	Message / content of the STW event	She said that she would resign$_{Message}$.
TOPIC	Topic of the STW event	Merkel addressed the theme of taxation$_{Topic}$.
EVIDENCE	Evidence for the message	The survey$_{Evidence}$ shows that ...
ADDRESSEE	Addressee of the STW event	Merkel spoke to the people$_{Addressee}$.
PARTICLE	Separated verb prefix or	Merkel schlug$_{Cue}$ vor$_{Particle}$ (proposed) ...
(PTC)	obligatory particle	Merkel stellt sich$_{Particle}$ vor (imagines herself) ...
MULTIWORD	multiword cue	I now give$_{Cue}$ you the floor$_{Multiword}$

Table 1: Overview over our schema for annotating events of **S**peech, **T**hought and **W**riting (STW).

Cue/Role	Freq.	Avg. length
CUE	7,706	1.1
SOURCE	4,663	1.7
MESSAGE	4,578	9.7
TOPIC	1,188	5.4
ADDRESSEE	717	3.2
PARTICLE	561	1.0
MEDIUM	321	3.2
EVIDENCE	151	4.3

Table 2: Statistics for our new dataset (CUE also includes multiword cues; 773 of the 7,706 cues are MULTIWORD cues).

destag – Open Data.[4] The data set includes 265 speeches from the German Bundestag, mostly from the 19th legislative term (2017-2021), given by 195 different speakers from 6 parties (CDU/CSU: 76, SPD: 57, AfD: 39, FDP: 33, The Left: 29, Greens: 26, non-attached: 4). The total size of the data is >200,000 tokens. For more detailed information on the data, sampling and annotation process, please refer to the datasheet.[5]

Annotation process The data was annotated by four student assistants from different fields in the humanities. The annotators received extensive training. During the annotation phase, weekly meetings were held where we discussed open questions and problematic cases.

To ease the detection of speech events, we started with a list of cue words extracted from the Corpus REDEWIEDERGABE (Brunner et al., 2020). We marked all lemma forms from the list in our data and instructed the annotators a) to verify whether each instance is a speech, thought and writing (henceforth: STW) event and, b) if true, to identify all of its arguments realised in the utterance. To increase recall, we asked the annotators to add new cue words to the list that were then included in the annotation. Table 2 shows the number of annotated cues and their roles in our corpus. Overall, we annotated more than 7,700 events of speech, thought or writing in the data.

Inter-annotator agreement We split the data into four samples that reflect the order of annotation. Table 4 shows the average percentage agreement of two coders for cue words and roles as the proportional token *overlap* between the annotated cues or roles. To augment this view, we also report a more lenient *binary* score which considers an annotation as correct if at least one token in the annotations overlaps and has been assigned the same label.[6] We can clearly see that inter-annotator agreement constantly improves with more training even after the third round of annotation.

Disagreements between the annotators Most questions during annotation concerned the class of thought events. Our guidelines follow Brunner et al. (2020) and define thought as "silent or inner speech which can be reproduced in the same way as verbalized speech". Brunner et al. (2020) conceptualise thought as "a conscious, analytical, cognitive process" and exclude descriptions of emotional and mood states or passages that are told from a strongly personal perspective. This definition, however, is hard to operationalise and there were many borderline cases that required discussion. We used our weekly meetings to decide which new cue words we would like to include. For more details, please refer to the annotation guidelines.

At the beginning of the annotation process, some annotators were eager to identify new cue words for thought events while others had a more conservative approach, considering only cues from our list. This is reflected in the high disagreement for sample 1. Sometimes new cues were included after one coder had already completed a docu-

[4] https://www.bundestag.de/services/opendata.

[5] The data, datasheet and annotation guidelines (in German) are available from our github repository: https://github.com/umanlp/spkatt.

[6] For more details on the scoring method, see (Marasovic and Frank, 2018).

A1 \ A2	ADDRESSEE	EVIDENCE	MEDIUM	MESSAGE	PARTICLE	SOURCE	TOPIC	NONE
ADDRESSEE	679	0	0	26	7	7	14	279
EVIDENCE	0	90	11	0	0	0	0	23
MEDIUM	0	64	109	17	0	5	18	245
MESSAGE	42	25	27	11,734	7	52	662	3,570
PARTICLE	0	0	0	8	101	0	1	46
SOURCE	22	15	8	106	1	2,244	22	623
TOPIC	4	3	0	214	0	0	574	194
NONE	310	116	48	3,530	91	407	335	0

Table 3: Confusion matrix (token level) for role annotations for the last two annotation samples.

ment, ignoring those cues, while the second coder included the new cues in the annotations. The confusion matrix (Table 3) shows that this is in fact the major source of disagreements: instances that were annotated by one annotator but not by the second coder (label NONE).

Other disagreements concern the distinction between MESSAGE and TOPIC (Example 3.1).

When distinguishing between TOPIC and MESSAGE, the annotators sometimes struggled to decide whether the speaker simply mentioned a certain topic or whether she also tried to convey a message. For instance, Example 3.1 may either be taken to mean that the addressee ("Sie" , 2Sg.formal) spoke about a democratic imposition (TOPIC) or that they said that something constituted a democratic imposition (MESSAGE).[7]

Ex. 3.1 (Topic vs. Message)

Sie haben von einer „demokratischen Zumutung" gesprochen.
You have spoken **of a "democratic imposition"**.

Another frequent class of disagreements includes MEDIUM vs. EVIDENCE. For illustration, see Example 3.2 where it is not clear whether the boldfaced text should be considered as the medium that transported the message or whether it should be interpretated as Evidence. More details on the distinction between those labels can be found in the annotation guidelines.

Ex. 3.2 (Medium vs. Evidence)

[...] die weltweite Stimmung mahnt uns, Erkämpftes zu erhalten [...]
[...] the global mood *urges us to preserve what we have fought for [...]*

In the next section we present an intrinsic evaluation of our new corpus by training a state of the art Semantic Role Labelling (SRL) system on our data.

[7] Based on the quotation signs used we think the latter interpretation is more likely to be correct but it's a subtle judgment.

	overlap		*binary*
Sample	**Cue**	**Roles**	**Roles**
Sample 1	69.07	64.53	67.88
Sample 2	81.19	67.04	72.60
Sample 3	81.95	72.11	76.90
Sample 4	82.84	73.81	77.63

Table 4: Pair-wise percentage agreement between the annotators on the four samples (*overlap:* proportional token overlap between A1 and A2; *binary:* at least one token in the cue/role span has been identified and assigned the same label).

4. Experiments

Task description The speaker attribution task consists in identifying the sources of speech, thought and writing in text, here, in parliamentary debates, and in linking the messages to their respective sources. The task can be decomposed into two subtasks.

1. Subtask 1: identify all trigger words that evoke a speech event
2. Subtask 2: for each speech event, identify all roles associated with this event (i.e., Source, Addressee, Message, Topic, Medium, Evidence)

The task setup is thus similar to Semantic Role Labelling (SRL), which allows us to utilize a state-of-the-art Semantic Role Labelling (SRL) system and train it on our data.

Model We adapt the SRL system of Conia and Navigli (2020) for our task. The system is language- and syntax-agnostic and jointly learns to predict the predicates, their senses and arguments (i.e., the roles of the speech event). In our setup, we use the predicate senses to encode whether a word form triggers a speech event or whether the same

		development set			test set		
	model	**prec**	**rec**	**F1**	**prec**	**rec**	**F1**
CUES	BERT-base	95.2$_{\pm 0.06}$	100.0$_{\pm 0.00}$	97.6$_{\pm 0.03}$	95.3$_{\pm 0.22}$	100.0$_{\pm 0.00}$	97.6$_{\pm 0.11}$
	BERT-large	95.2$_{\pm 0.26}$	100.0$_{\pm 0.00}$	97.6$_{\pm 0.14}$	95.9$_{\pm 0.01}$	100.0$_{\pm 0.00}$	97.9$_{\pm 0.00}$
ROLES	BERT-base	81.0$_{\pm 0.65}$	79.1$_{\pm 1.43}$	80.0$_{\pm 0.71}$	78.8$_{\pm 1.62}$	82.0$_{\pm 0.67}$	80.4$_{\pm 0.63}$
	BERT-large	81.3$_{\pm 1.18}$	78.6$_{\pm 0.16}$	79.9$_{\pm 0.0.65}$	79.9$_{\pm 0.30}$	82.1$_{\pm 0.10}$	81.0$_{\pm 0.20}$

Table 5: Avg. results (token overlap) for cue words and role prediction (dev/test) and standard deviation over three runs.

word is used with a different reading (marked as NONE). Example 2 shows two senses of the word *heißen* where only the first evokes a speech event.

(1) [In der Allgemeinen Erklärung der Menschenrechte von 1948]$_{Medium}$ heißt es: "[Jeder Mensch hat den Anspruch auf eine Staatsangehörigkeit]$_{Message}$."
[The 1948 Universal Declaration of Human Rights]$_{Medium}$ states: "[Every human being has the right to a nationality]$_{Message}$."

(2) Das heißt$_{NONE}$, wir nehmen das sehr genau unter die Lupe.
That means$_{NONE}$ we are taking a very close look at it.

The SRL model of Conia and Navigli (2020) combines a predicate-aware word encoder with a predicate-argument encoder. The first component yields contextualized word representations with respect to the predicate of the sentence, while the second encoder learns predicate-aware argument representations. To identify the predicate candidates, the system uses a dictionary of lemma forms that are then disambiguated, based on the encoded representations. We extract the dictionary from the training data. This means that predicates in the development and test set that have not been seen during training are not considered by the system and will get penalised in the evaluation. In our experiments, however, this is not a problem as our training data is large enough to provide sufficient coverage.

We initialize the SRL system with the pretrained gbert-base and gbert-large[8] language models (Chan et al., 2020) and select the best fine-tuned model on the dev set.[9] For that, we split our data into training, dev and test sets with 9,298/927/3,067 sentences.[10] This amounts to 178/18/72 different speeches in each set, with 5,536 (train), 515 (dev) and 3,646 (test) annotated events of speech, thought and writing.

Evaluation metric The evaluation of system performance uses the familiar Precision, Recall and F1 metrics. The first task of the system is to disambiguate whether a given cue candidate, identified by the lemma dictionary, does evoke a speech event in a specific context or not (label: NONE).

Both cue and role labels can cover more than one token and therefore are represented as sets of (possibly discontinuous) tokens. The annotation scheme assumes that a given set of tokens can bear at most one cue annotation, that is, it can evoke at most one instance of speech, throught or writing. For roles this is not true: a set of tokens could bear multiple role labels, usually in relation to different cues. According to our annotation guidelines, roles are dependent on cues and so system roles can match gold roles only if they are related to the same cue. To avoid a many-to-many mapping between multiword cues in the gold data and system output, we only consider the head of a MULTIWORD construction as a cue and evaluate the other multiword components as part of the roles (MULTIWORD).

In line with this, the evaluation first checks how system cues and gold cues align. In doing so the scorer matches at most one system cue to at most one gold cue and the same in the other direction. System cues that cannot be aligned to gold cues produce false positives, including for their associated roles. In symmetric fashion, gold cues that cannot be aligned to a system cue result in false negatives.

For both cues and roles, alignment requires non-zero overlap with the tokens covered by a label of the same type on the other side. Each component token of aligned labels is counted as a true or false positive, or as a false negative. This means that longer spans contribute more to the overall score than shorter labels.

To illustrate the evaluation of multiword cues, consider the following example.

[8] https://huggingface.co/deepset/gbert-base

[9] To ensure replicability, we will release the trained model and configuration files together with the train/dev/test splits.

[10] We use spacy for sentence segmentation which results in segments on the clause level, with an average size of around 16 tokens/clause.

Role	prec	rec	F1
SOURCE	82.4± 1.22	85.7± 0.38	84.0± 0.75
MESSAGE	81.3± 0.81	84.5± 0.29	82.8± 0.44
ADDRESSEE	81.5± 0.95	75.7± 2.92	78.5± 1.86
TOPIC	69.0± 2.27	70.1± 1.67	69.6± 1.78
EVIDENCE	84.4± 1.15	64.7± 4.82	73.2± 3.36
MEDIUM	65.2± 5.88	68.3± 0.59	66.6± 3.37
PARTICLE	81.9± 2.13	82.9± 1.38	82.4± 1.51
MULTIWORD	61.8± 2.11	62.7± 3.73	62.2± 2.43

Table 6: Results for individual roles on the test set, averaged over three runs, and standard deviation.

Ex. 4.1 (Evaluation of MULTIWORD cues)
eine Rede$_{Multiword}$ im Bundestag halten$_{Cue}$ giving$_{Cue}$ a speech$_{Multiword}$ in parliament

We consider the cue as a true positive if the system has correctly identified "halten" (give) as a cue. We then evaluate all roles attached to the cue, including the MULTIWORD component "Rede" (speech).

Evaluation results Table 5 shows results for the fine-tuned models on the development and test set. The SRL system yields high scores for the detection of cue words and their roles on both sets. The development set seems to include more difficult instances, as shown by the slightly higher scores for roles for the test set. Overall, the results are satisfactory, with reasonably high precision and recall for both, triggers and roles.

Looking at the results for individual roles (Table 6), we can see that the model struggles with the same roles that our human coders found difficult, i.e., EVIDENCE, MEDIUM and TOPIC (see §3, Table 3). It is, however, not clear whether this is due to the inherent ambiguity of the labels or whether it simply reflects the frequency of the labels in the training data, as EVIDENCE and MEDIUM are the labels with the lowest number of instances in our data (see Table 2). Identifying the components of MULTIWORD cues is also hard for the model, given that they are mostly discontinuous and are not that frequent.[11]

5. Exploring speech events in 20 years of parliamentary debates

We now showcase how our new resources can contribute to the analysis of political text.

5.1. Corpus creation and preprocessing

For our case study, we use a large corpus of parliamentary debates from the German Bundestag, ranging from 2002 to 2023. The first part of the data comes from the German subset of the ParlSpeech V2 corpus (Rauh and Schwalbach, 2020) which includes parliamentary debates from the German Bundestag from 1991 to 2018.[12]

We augment this dataset with newer speeches downloaded from the open data service of the German Bundestag, covering a time range from January 2019 to September 2023.[13] The newer data is available in an xml format where one file includes the speeches for all agenda items that have been discussed on a particular day. We split the data into individual speeches and augment the text with metadata including the speakers' names, their party affiliation and the date of the speech.

Preprocessing The text in our corpus is originally spoken, even though many spoken language phenomena such as filled pauses were already removed during transcription by the keeper of the minutes. As we noticed that the German spaCy models, which are trained on newspaper text, are not very good at handling parentheses, we removed frequently occurring parenthetical clauses from the unlabelled data before sentence splitting in order to obtain coherent and connected utterances.

Ex. 5.1 (Parenthetical clauses)

"Als nationale Antwort auf PISA ist das von Ihnen vorgeschlagene Schulbauprogramm – mehr ist es nicht – ungeeignet."
As a national response to PISA, your proposed school building program – that's all it is – is unsuitable.

This removal was straightforward since parenthetical clauses were clearly marked off by the use of double hyphens. After preprocessing, our data has a size of over 73 mio. tokens.

5.2. Identifying epistemological bias in parliamentary debates

In this section, we investigate epistemological bias (Recasens et al., 2013), i.e., how political actors frame their messages when talking about political events. More specifically, we want to know which propositions are framed as facts or common ground by the speaker rather than being presented as personal opinions.

[11] Our corpus includes 470/35/268 MULTIWORD cues in the training/dev/test set.

[12] After preprocessing the data, we noticed that the cabinet protocols from 1991 to 2001 do not seem to be complete. We therefore decided to restrict our analysis to the years from 2002 to 2023.

[13] https://www.bundestag.de/services/opendata

Parties in government	**Election year**	**Leg. term**	**Parties in the German Bundestag**					
			AfD	**CDU/CSU**	**FDP**	**GREENS**	**LEFT**	**SPD**
SPD+Greens	2002	15	0	129,668	42,397	53,104	9,172	110,826
CDU+SPD	2005	16	0	144,136	80,360	87,484	70,797	155,067
CDU+FDP	2009	17	0	201,485	99,386	88,496	82,927	158,353
CDU+SPD	2013	18	0	182,443	–	89,663	75,998	132,673
CDU+SPD	2017	19	63,739	188,516	62,815	63,445	52,333	121,519
SPD+FDP+Greens	2021	20	29,633	87,575	41,204	44,997	20,184	79,587
		total	93,372	1,108,444	393,970	478,683	351,433	900,967
		total	194,140	1,816,078	634,861	761,916	606,545	1,435,419
SPD+Greens	2002	15	0	0.661	0.689	0.678	0.629	0.656
CDU+SPD	2005	16	0	0.625	0.671	0.693	0.638	0.655
CDU+FDP	2009	17	0	0.600	0.600	0.634	0.577	0.659
CDU+SPD	2013	18	0	0.570	–	0.634	0.580	0.607
CDU+SPD	2017	19	0.491	0.587	0.563	0.577	0.514	0.577
SPD+FDP+Greens	2021	20	0.461	0.592	0.543	0.499	0.472	0.523
		avg	0.476	0.633	0.654	0.642	0.606	0.655

Table 7: Number of speech events (upper part of the table) and frequencies normalised by no. of tokens (below) per party for each legislative term in our data (2002–2023). The first column encodes the coalition in government (CDU stands for the conservative union of CDU/CSU).

Parties in government	Election year	Leg. term	Parties in the German Bundestag					
			AfD	CDU/CSU	FDP	GREENS	LEFT	SPD
SPD/GREENS	2002	15	–	6.9	7.5	7.8	7.0	7.9
CDU/SPD	2005	16	–	7.4	7.5	7.1	6.4	7.6
CDU/FDP	2009	17	–	7.6	8.2	7.2	6.6	7.1
CDU/SPD	2013	18	–	8.2	–	7.3	7.0	8.0
CDU/SPD	2017	19	8.2	7.9	7.8	7.6	6.9	8.4
SPD/FDP/GREENS	2021	20	9.0	7.6	9.3	9.8	7.3	9.5

Table 8: Proportion of factives used by the different parties. Underlined numbers indicate that the respective party was in government at this time and blue indicates that the party was part of the opposition.

Step I: Attributing messages to speakers In the first step, we want to find out who says what to whom, and predict speech events (including thought and writing) for all speeches in our corpus. To this end, we extract all trigger words and their roles. Table 7 shows the number of speech events for each party and legislative term in our data. The far-right AfD was first elected in the German Bundestag in 2017 (19th legislative term), and the liberal party FDP failed to clear the 5% hurdle in the elections in 2013 (18th legislative term).

Step II: Accepted fact or opinion? We now want to identify propositions that are framed as a fact by the speaker. Example 5.2 below presents its statement as a generally accepted fact that is part of the common ground, while the speaker in the second example explicitly marks the proposition as a personal opinion.

Ex. 5.2 For 100 days, *everyone in this room*$_{Source}$ has also known **that we live in a new reality and that this reality remains, the reality of one state invading another state in Europe for no reason**$_{Message}$. *(Faber, FDP, 2022-03-06)*

Ex. 5.3 *We*$_{Source}$ are of the opinion **that the violation of territorial integrity must not be and that Ukraine must regain access to its entire territory**$_{Message}$. *(Merkel, CDU/CSU, 2019-06-26)*

To identify factivity in our data, we created a lexicon of factive speech event triggers, based on and extending the English verb lists by Hooper (1975). Our lexicon non only includes verbs but also nouns and multiword expressions.

German factivity lexicon Our lexicon contains 84 entries of factives (21 verbs and 63 multi-word expressions (MWEs)). Please note that we extracted the entries as lemma forms in the order

they appeared in the text. We did not normalise word order, meaning that multi-word expressions such as "enttäuscht zeigen" and "zeigen enttäuscht" (engl.: show disappointed) are included as two separate entries. Discounting word order variation, the lexicon includes 63 factives.[14]

We can now use this lexicon to identify factive propositions in parliamentary debates from the German Bundestag, together with their sources and messages and all other roles filled for the respective utterances.

Parties in government and opposition show different use of factives Table 8 gives an overview of the proportion of factives in the speeches for each party and legislative term. One striking difference emerges: members of the government (underlined numbers) make more frequent use of factives than members of opposition parties (blue numbers). After controlling for normal distribution (using the Shapiro-Wilk test) and for homogeneity of variance (Levine test), we applied a one-way ANOVA and found that the difference in the use of factives between the two groups is statistically significant ($p < 0.001$).

To investigate the observed differences between the rhetorical behaviour of government and opposition parties, we use the extracted roles for the speech events that are framed as facts by the speakers. We focus on the roles SOURCE, MESSAGE and TOPIC as all three are used frequently in our data and can be predicted by the classifier with F1 scores in the range of 76% to 89%. We then use sentence embeddings (Reimers and Gurevych, 2019) to encode the texts for each role and apply the Fast Clustering algorithm provided in the sentence transformers package.[15] We experiment with similarity thresholds $\theta = [0.75, 0.8]$ and minimum cluster sizes $s = [25, 100]$. As a result, we get role-specific clusters for the sources, messages and topics of factive speech events.

Our goal is to identify topical clusters in the debates that fill a certain role (either SOURCE MESSAGE or TOPIC) and to identify patterns of usage for parties that are in government and parties in the opposition. For that, we first compute how often each custer is used by a specific party, then normalise the raw frequencies by the number of speech events, and finally compute a confidence interval for each cluster. We use this confidence interval to identify which clusters are used significantly more (or less) often by a specific party or group, i.e., with normalised frequencies outside of the confidence intervals.

[14]The lexicon is available from our github repository: https://github.com/umanlp/spkatt.

[15]https://www.sbert.net/

Table 9 shows the number of clusters for two different settings, (1) more coarse-grained clusters with a size of 100 and more fine-grained clusters with a size of 25. For each party, the table shows how often members of this party use a certain cluster more/less often than the other parties. We can see some variation for the different cluster sizes, however, when looking at the ranking of the parties, we find that for both the coarse-grained and the fine-grained clusters, the far-right AfD deviates most often, followed by the left-wing party The Left.

Qualitative analysis To gain more insight into this matter, we add a qualitative analysis where we extract the role fillers for the AfD and The Left for those clusters where the respective parties most deviate from the mean. One of the clusters for messages overused by the far-right AfD includes negative statements, as shown in Example 3 below.

(3) MESSAGE, Cluster 3 (547 instances)
   a. dass es nicht geht
      that it doesn't work
   b. dass es damit nicht getan ist
      that this is not enough
   c. dass es nicht möglich ist
      that it is not possible
   d. dass das nicht zulässig ist
      that this is not allowed

The next example shows a cluster for sources that refer to the people and that has been underused by the left-wing party The Left. This comes a bit as a surprise, given that The Left has received a high score of 6.9 (out of 10) for *people-centrism* in the Populism and Political Parties Expert Survey (POPPA) (Meijers and Zaslove, 2021), and warrants further investigation.

(4) SOURCE, Cluster 7 (164 instances)
   a. Die Bürger in unserem Land
      The citizens of our country
   b. die Menschen auf dem Land
      the people in the countryside
   c. die Bürger draußen
      the citizens outside
   d. die Mehrzahl der Bevölkerung
      the majority of the population

For topic, we find that the liberal FDP makes a more frequent use of Cluster 6 which includes mentions of the state budget (Example 5).

(5) SOURCE, Cluster 7 (164 instances)
   a. im Budget
      in the budget
   b. den Haushalt 2024
      the budget 2024

Cluster size	Threshold	Role	# clusters	AfD	CDU	FDP	GREENS	LEFT	SPD
100	0.75	MESSAGE	261	4	1	0	0	3	1
100	0.75	SOURCE	38	4	0	2	0	1	1
100	0.75	TOPIC	4	2	0	0	0	0	0
			total	10	1	2	0	4	2
25	0.80	MESSAGE	910	5	1	0	0	5	1
25	0.80	SOURCE	136	4	1	2	1	5	1
25	0.80	TOPIC	35	4	2	4	4	1	2
			total	13	4	6	5	11	4

Table 9: Number of extracted clusters for the speech event roles SOURCE, MESSAGE and TOPIC co-ocurring with factive speech event triggers. The last six columns show the number of clusters where parties over-/underuse a specific cluster, compared to the sample mean (confidence threshold=0.99).

c. beim Haushalt des Bundesministers
at the budget of the Federal Minister

d. mehreren Stellen im Haushalt
several positions in the budget

These examples are meant to illustrate how speaker attribution can be used to investigate what political actors talk about and how they frame their messages.

6. Conclusions

We presented GePaDe$_{SpkAtt}$, a new resource for speaker attribution in German parliamentary debates, with more than 7,700 manually annotated speech events. We show that our annotations can be predicted with high F1 scores, yielding over 97% F1 for cue words and over 80% F1 for roles. We then applied our classifier to predict speech events in a large corpus of parliamentary debates and explored the use of factives in political rhetoric, showing that members of the government make more frequent use of factives in their speeches than members of the opposition.

All resources described in the paper are made available to the research community. In future work, we plan to use our resources to investigate framing in parliamentary debates.

7. Limitations

One important limitation of using our classifier for political text analysis is that we have to rely on the automatic predictions of the model, knowing that these predictions do include a certain amount of errors. This is not problematic as long as the errors are random, meaning they are evenly distributed across speakers and parties. While we do not expect a huge variation in the use of *verba dicendi* in parliamentary debates from the last two decades, it is however conceivable that we find systematic differences when looking at larger time spans, or that there might be differences concerning the communicative behaviour of the different parties. To test and quantify this potential error is by no means straightforward but has to be addressed in future work.

Acknowledgements

The work presented in this paper was funded in part by the German Research Foundation (DFG)under the UNCOVER project (RE3536/3-1). We would also like to thank the German Society for Computational Linguistics & Language Technology (GSCL) for their generous financial support. Finally, we would like to thank our student annotators for their dedicated work.

8. Bibliographical References

Mariana S. C. Almeida, Miguel B. Almeida, and André F. T. Martins. 2014. A joint model for quotation attribution and coreference resolution. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 39–48.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Annelen Brunner, Stefan Engelberg, Fotis Jannidis, Ngoc Duyen Tanja Tu, and Lukas Weimer. 2020. Corpus REDEWIEDERGABE. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC),*

May 11-16, 2020, Palais du Pharo, Marseille, France, pages 803 – 812, Paris. European Language Resources Association.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Identifying sources of opinions with conditional random fields and extraction patterns. In *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, HLT/EMNLP 2005, pages 355–362.

Simone Conia and Roberto Navigli. 2020. Bridging the gap in multilingual semantic role labeling: a language-agnostic approach. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1396–1410, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tillmann Dönicke, Hanna Varachkina, Anna Mareike Weimer, Luisa Gödeke, Florian Barth, Benjamin Gittel, Anke Holler, and Caroline Sporleder. 2022. Modelling Speaker Attribution in Narrative Texts With Biased and Bias-Adjustable Neural Networks. *Frontiers in Artificial Intelligence*, 4:725321.

David K. Elson and Kathleen R. McKeown. 2010. Automatic attribution of quoted speech in literary narrative. In *The Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI 2010.

Susanne Fertmann. 2016. Using speaker identification to improve coreference resolution in literary narratives. Master's thesis, Computational Linguistics.

Yulia Grishina and Manfred Stede. 2017. Multi-source projection of coreference chains: assessing strategies and testing opportunities. In *The 2nd Coreference Resolution Beyond OntoNotes Workshop*, CORBON-2017.

Hua He, Denilson Barbosa, and Grzegorz Kondrak. 2013. Identification of speakers in novels. In *The 51st Annual Meeting of the Association for Computational Linguistics*, ACL 2013, pages 1312–1320.

Joan B. Hooper. 1975. On assertive predicates. *Syntax and Semantics*, 4:91–124.

Richard Johansson and Alessandro Moschitti. 2013. Relational features in fine-grained opinion analysis. *Computational Linguistics*, 39(3):473–509.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ralf Krestel, Sabine Bergler, and René Witte. 2008. Minding the source: Automatic tagging of reported speech in newspaper articles. In *The International Conference on Language Resources and Evaluation*, LREC 2008.

Markus Krug, Frank Puppe, Isabella Reger, Lukas Weimer, Luisa Macharowsky, and Stephan Feldhaus. 2018. *Description of a Corpus of Character References in German Novels – DROC [Deutsches ROman Corpus]. DARIAH-DE Working Papers. Göttingen: DARIAH-DE.*

Ana Marasovic and Anette Frank. 2018. SRL4ORL: improving opinion role labeling using multi-task learning with semantic role labeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 583–594. Association for Computational Linguistics.

Maurits J. Meijers and Andrej Zaslove. 2021. Measuring populism in political parties: Appraisal of a new approach. *Comparative Political Studies*, 54(2):372–407.

Grace Muzny, Angel X. Chang, Michael Fang, and Dan Jurafsky. 2017. A two-stage sieve approach for quote attribution. In *The 15th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2017, pages 460–470.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Silvia Pareti. 2015. *Attribution: a computational approach*. Ph.D. thesis, University of Edinburgh, UK.

Silvia Pareti, Timothy O'Keefe, Ioannis Konstas, James R. Curran, and Irena Koprinska. 2013. Automatically detecting and attributing indirect

quotations. In *The 2013 Conference on Empirical Methods in Natural Language Processing*, EMNLP 2013, pages 989–999.

Bruno Pouliquen, Ralf Steinberger, and Clive Best. 2007. Automatic detection of quotations in multilingual news. In *The International Conference on Recent Advances in Natural Language Processing*, RANLP 2007, pages 487–492.

Christian Rauh and Jan Schwalbach. 2020. The ParlSpeech V2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies. Available from Harvard Dataverse.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Josef Ruppenhofer, Caroline Sporleder, and Fabian Shirokov. 2010. Speaker attribution in cabinet protocols. In *The Seventh conference on International Language Resources and Evaluation*, LREC 2010.

Josef Ruppenhofer, Julia Maria Struß, and Michael Wiegand. 2016. Overview of the IGGSA 2016 Shared Task on Source and Target Extraction from Political Speeches. In Josef Ruppenhofer, Julia Maria Struš, and Michael Wiegand, editors, *IGGSA Shared Task on Source and Target Extraction from Political Speeches*, pages 1 – 9. Ruhr-Universität Bochum, Bochum.

Christian Scheible, Roman Klinger, and Sebastian Padó. 2016. Model architectures for quotation detection. In *The 54th Annual Meeting of the Association for Computational Linguistics*, ACL 2016.

Michael Wiegand and Dietrich Klakow. 2012. Generalization methods for in-domain and cross-domain opinion holder extraction. In *The 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL 2012, pages 325–335.